Full multicondition training for robust i-vector based speaker recognition

Dayana Ribas[1], *Emmanuel Vincent*[2], *José Ramón Calvo*[1]

[1]Advanced Technologies Application Center (CENATAV), Habana, Cuba
[2]Inria, Villers-lès-Nancy, F-54600, France
{dribas,jcalvo}@cenatav.co.cu, emmanuel.vincent@inria.fr

Abstract

Multicondition training (MCT) is an established technique to handle noisy and reverberant conditions. Previous works in the field of i-vector based speaker recognition have applied MCT to linear discriminant analysis (LDA) and probabilistic LDA (PLDA), but not to the universal background model (UBM) and the total variability (T) matrix, arguing that this would be too much time consuming due to the increase of the size of the training set by the number of noise and reverberation conditions. In this paper, we propose a *full MCT* approach which consists of applying MCT in all stages of training, including the UBM and the T matrix, while keeping the size of the training set fixed. Experiments in highly nonstationary noise conditions show a decrease of the equal error rate (EER) to 14.16% compared to 17.90% for clean training and 18.08% for MCT of LDA and PLDA only. We also evaluate the impact of state-of-the-art multichannel speech enhancement and show further reduction of the EER down to 10.47%.

Index Terms: speaker recognition, robustness, multicondition training, UBM, speech enhancement.

1. Introduction

A current challenge in the field of speaker recognition is to migrate automatic systems developed in the lab to real world environments. The distortion of speech by environmental noise and reverberation jointly with channel mismatch provokes a variability that degrades considerably the high accuracy reached in the lab. During the last decade, efforts have focused on the development of the speaker recognition framework based on factor analysis [1–4]. This approach is capable of managing the linear variability due to the different channels and test sessions, however the nonlinear variability due to noise and reverberation cannot be well managed in this way.

Several robust methods have been developed for automatic speech recognition (ASR) [5–7] and speaker recognition [8] that seek to compensate for speech distortion in the input features, in the model parameters, or both. MCT, also called multistyle training, is an effective technique to handle speech signals acquired in different conditions. In theory the optimal recognition performance is obtained in matched conditions, that is when training the system on a dataset acquired in the same conditions as the test utterance. Unfortunately, the test conditions are often unknown [9]. MCT handles this issue by training the system on a dataset involving as diverse conditions as possible, such as different speaking styles [10], sampling rates [11], languages [12], or speaker ages [13], among others. This generally contributes to improving performance, even if the precise test conditions were never observed before [14].

The MCT concept has shown to improve the performance of speaker recognition systems in noisy environments both in the previous Gaussian mixture model (GMM)-UBM framework [15] and in the current i-vector-PLDA framework [16–19]. In the latter case, MCT has been applied to the i-vector projection matrix based on LDA and to the final scoring based on PLDA, both for enrollment and test. There is no report, however, about the effect of MCT on the T matrix and the UBM. The main reason for this is that the usual methodology consists of replicating the training dataset for each condition. Even though training is performed offline and it does not affect the computation time at test, this increases considerably the training size and makes it time consuming.

The combination of MCT with front-end speech enhancement generally improves performance in the context of ASR [6], but it is believed to worsen results in the context of speaker recognition [20]. This belief may be due to the historical focus on single-channel speech enhancement techniques, some of which strongly distort speech. However, many applications now offer multichannel speech input and can benefit from powerful multichannel enhancement techniques.

In this work, we propose a *full MCT* approach for noise robust speaker recognition which consists of applying MCT to all stages of the i-vector-PLDA framework, while keeping the size of the training set fixed. We evaluate this approach through a speaker verification experiment on NIST-SRE data [21] in the line of [18] and we assess the benefit of multichannel speech enhancement, alone and combined with MCT.

In Section 2, we review past studies on MCT in the i-vector-PLDA framework and introduce the proposed approach. Section 3 describes the experimental setup. Finally Sections 4 and 5 present the results, discussion, and conclusions of the paper.

2. Partial vs. full MCT

2.1. Acoustic distortion

The ubiquity of acoustic distortions in speaker recognition applications is a fact. Applications such as user authentication in telephone cabins, door access control, ATM bank transactions, forensic investigation, or home automation, among others, are affected by reverberation, additive noise, or both. This implies a mismatch with the model parameters, which are usually trained on clean speech, which decreases the recognition performance.

The type and level of distortion are the result of several variables. Indoor recordings are usually affected by reverberation, which depends on the size and the features of the room, as well as the position of the speaker in the room. Both indoor and outdoor recordings may also be affected by noise, which depends on the other acoustic sources in the scenario. The magnitude of the speech signal with respect to the environmental distortion

is inversely proportional to the distance from the speaker to the microphone [22].

In such applications, each test utterance may exhibit different types and levels of distortion depending on the conditions at the time of recording. These conditions are generally unknown, hence the attractiveness of MCT.

2.2. Partial MCT in the i-vector-PLDA framework

The i-vector-PLDA framework for speaker verification involves four processing stages [3,4]: UBM, T matrix, LDA, and PLDA. Several authors have applied the MCT concept to this framework and reported positive results.

Some works have focused on training the LDA matrix from a multicondition set or averaging the between-class and within-class covariance matrices learned from speech in different conditions. In [23], the LDA matrix was trained using pooled telephonic and microphonic speech. In [24–26] the authors studied different ways of estimating and averaging the between-class and within-class covariance matrices.

Other works have been directed to the training of the PLDA model parameters. In [27], the PLDA model is trained using pooled clean and noisy speech, while in [28] the multicondition set involves telephonic and microphonic speech. In [29], three PLDA models are created using speech from three different channels and the resulting three scores are fused. The authors in [16] proposed a similar approach based on training a collection of systems tuned to specific conditions and on fusing their scores. In [17], the authors proposed a variant of PLDA where the channel space is estimated only from the data of the corresponding channel, while the speaker space is estimated from all available data.

The impact of multicondition enrollment data has also been studied. The results in [18] indicate that, when PLDA is trained on a multicondition set, the use of a multicondition enrollment set yields significant performance improvement compared to a clean enrollment set. Finally, the study in [19] has assessed single-channel feature-domain noise compensation methods in combination with MCT.

In the following, we refer to the application of MCT in the LDA and PLDA stages as *partial MCT* [18]. To the best of our knowledge, none of the above works has performed MCT over the other training stages, namely the UBM and the T matrix. This was motivated by the fact that the computation of the UBM and the T matrix would be too much time consuming [18] when the training data are replicated for every noise condition.

2.3. Proposed full MCT

By contrast, we propose to apply MCT to all training stages and we call this approach *full MCT*. We also optionally consider a multicondition enrollment set. In order to deal with the computational cost issue, we keep the size of the multicondition training and enrollment datasets equal to those of the clean training and enrollment datasets. Specifically, the multicondition datasets consist of the same amount of speech as the clean datasets, but each utterance exhibits one random reverberation and noise condition. The evaluation dataset also includes different reverberation and noise conditions, which are considered to be unknown. The overall computational cost of multicondition training and enrollment is therefore equal to that of clean training and enrollment and it does not affect the computational cost of test.

This MCT configuration has been known to provide significant performance improvement in the field of ASR compared to clean training, while the duplication of the training data often provides only moderate additional improvement in that context (see, e.g., [30, 31]).

In order to further improve the potential of full MCT, we propose to exploit it in combination with multichannel noise reduction. To do so, we apply the same noise reduction technique to the enrollment and test datasets and we train the system by MCT on reverberated noiseless speech.

3. Experimental setup

A majority of noise-robust speaker verification studies have relied on simulated data, due to the lack of recordings of human speakers in real noise scenarios [15,19,20,27]. Recently, NIST-SRE 2012 made a step towards addressing this need in single-channel conditions, but there is still no such dataset in multichannel conditions. Yet, as mentioned earlier, many applications now offer multichannel speech input and can benefit from powerful multichannel enhancement techniques. In the following, we carry out a series of text-independent speaker verification experiments by using noise and reverberation from Track 1 of the 2nd CHIME Challenge [30] to corrupt clean speech from NIST-SRE. These data were recorded in a real domestic environment and they stand out from other multichannel datasets by the attention brought to the realism of the sound scenes, which include multiple, highly nonstationary noise sources, and were not rescaled to alter the signal-to-noise ratio (SNR). They received significant attention in the robust ASR community [31].

3.1. Speech corpus

The experiments were carried out using male conversations in English. For the training stage (UBM, T matrix, LDA, PLDA), 3285 speech signals of 262 speakers from NIST-SRE 2004 and 2005 were used. For the evaluation stage, the *short2* and *short3* datasets of NIST-SRE 2008 were employed, namely 470 speech signals for enrollment and 671 speech signals for test. A total of 6615 verifications were performed on the *det 7* condition of NIST-SRE, including 439 targets. All signals are 16-bit telephonic signals sampled at 8 kHz. Each signal has around 5 min duration, with around 1 min of useful speech.

3.2. Noise and reverberation conditions

A set of 121 two-channel room impulse responses (RIRs) were measured in a domestic living room with a reverberation time (RT) of 0.3 seconds. Several hours of background noise were also recorded in that room, including voices, TV, game console, cutlery sounds, footsteps, etc [30]. Each *clean* speech signal was convolved with one randomly chosen RIR, resulting in a two-channel *reverberated* signal with the same duration as the clean signal. This signal was then mixed with one randomly chosen segment of background noise, resulting in a two-channel *noisy* signal. Each utterance is therefore available under three forms: clean (original NIST), reverberated (without noise), and noisy (with reverberation and noise).

Different background noises were used for training and for enrollment and test. The resulting SNR was computed as the ratio of the energy of the two-channel reverberated speech signal and the two-channel noise signal and ranged from about -10 to +20 dB with an average of 6.1 dB. Note that the local SNR may differ from the global SNR and that both the type and the level of noise vary significantly with the SNR [30], hence the resulting data include multiple noise conditions. These conditions were assumed to be unknown both in the training stage

and in the enrollment and test stages.

3.3. Speech enhancement

Speech enhancement was applied to the enrollment and test sets using the Flexible Audio Source Separation Toolbox (FASST) [32], which has shown state-of-the-art performance on the CHiME data [33]. This toolbox leverages both single- and multichannel characteristics of speech and noise in order to perform enhancement. Target speech is considered as a single source and background noise is modeled as the sum of two sources. The short-term power spectrum of each source is modeled as a linear combination of 32 basis spectra via nonnegative matrix factorization (NMF), while its spatial position and spatial width are modeled by its full-rank spatial covariance matrix at each frequency. The basis spectra and the spatial covariance matrices of target speech are learned from the reverberated training set, while those of the background noise sources are trained on the 5 s immediately before and after each speech activity interval in the test utterance. For detailed settings, see [33].

3.4. Speaker verification system

The speaker verification system used follows the state of the art i-vector-PLDA framework [3]. It uses a Mel frequency cepstral coefficient (MFCC) front-end, computed over Hamming-windowed frames with 20 ms size and 10 ms shift. The MFCCs were obtained using a Mel filterbank of 24 channels, followed by a transformation to the cepstral domain, keeping 19 coefficients and computing the log-energy. The first and second derivatives of the cepstral coefficients were added, followed by frame selection using voice activity detection (VAD) and cepstral mean and variance normalization (CMVN) [14].

The UBM consists of 512 Gaussians. For the i-vector extraction, a T matrix of dimension 400 and an LDA matrix of dimension 330 are used. The i-vectors are projected with LDA, after each i-vector is centered, whitened and length-normalized. Classification relies either on plain LDA or on Gaussian PLDA, as explained in [16].

3.5. Training, enrollment, and test configurations

We compared three different training techniques:

- without MCT: training on clean (original NIST) data,
- partial MCT [18]: training the UBM and the T matrix on clean data, and LDA and PLDA on multicondition data,
- full MCT (proposed): training the UBM, the T matrix, LDA, and PLDA on multicondition data.

For partial MCT and full MCT, either reverberated data or the noisy data were used as multicondition training data, resulting in five training configurations in total. Enrollment and test were performed on three different versions of the data, namely clean, noisy, or enhanced.

Note that the speech enhancement method seeks to reduce noise but not reverberation. Hence, the distribution of noisy test data matches the one of noisy training data, while the distribution of enhanced test data ideally matches the one of reverberated (not clean) training data.

4. Results and Discussion

This section presents the results of the above experiments. All results are expressed in terms of equal error rate (EER) and minimum value of the NIST detection cost function (mDCF) [21].

Note that we report the results for both classification approaches: LDA and PLDA. Indeed, although the results of LDA are not as good as those of PLDA, LDA can be more convenient for those applications where speed matters more than accuracy.

4.1. Performance on clean data

Table 1 presents the results for the speaker recognition system trained and tested with clean (original NIST) signals. This provides a bound on the performance achievable in noisy and reverberated conditions, which is discussed hereafter.

Table 1: *Speaker recognition results on clean data.*

Clean training, enrollment, and test data			
LDA		**PLDA**	
EER	**mDCF**	**EER**	**mDCF**
4.91	0.022	3.19	0.018

4.2. Performance on noisy or enhanced data

Tables 2 and 3 show the performance for various MCT configurations on noisy or enhanced test data.

4.2.1. Evaluation using reverberated and noisy speech in the enrollment dataset

Table 2 shows the results with noisy or enhanced enrollment data. The best result for each test configuration is highlighted in gray. It can be seen that the proposed full MCT approach outperforms the previous partial MCT approach in [18] for both test sets: noisy and enhanced. Full MCT also outperforms clean training in a majority of cases. It can also be seen that multichannel speech enhancement in the test set provides improvement for all training approaches and training data.

Overall, the best results for the noisy test set are obtained with full MCT on the noisy training set and the best results for the enhanced test set are obtained with full MCT on the reverberated training set. These results support the theory that increasing the level of matching between training and test conditions generally results in increased system performance [9].

4.2.2. Evaluation using clean speech in the enrollment dataset

Table 3 shows the performance of the same MCT approaches as in Table 2, however in this case the clean (original NIST) signals are used for enrollment.

As before, full MCT always outperforms partial MCT and clean training, provided that the best matched training set (noisy or reverberated) is chosen. Also, multichannel speech enhancement decreases both the EER and the mDCF for all training approaches and training data.

4.2.3. Summary results

Comparing Tables 2 and 3, we conclude that, independently of the acoustical conditions used for enrollment, the extension of the MCT concept to all training stages can provide better accuracy than MCT applied to LDA and PLDA only.

Using PLDA classification, the best EER across the two tables is equal to 14.16% for full MCT (with noisy enrollment) compared to 17.90% for clean training (with clean enrollment) and 18.08% for partial MCT (with noisy enrollment). Multichannel speech enhancement further reduces the EER down to 10.47% (with enhanced enrollment).

Table 2: *Speaker recognition results using noisy or enhanced enrollment and test data.*

Type of MCT	Training of UBM and T matrix	Training of LDA and PLDA	Noisy enrollment and test data				Enhanced enrollment and test data			
			LDA		PLDA		LDA		PLDA	
			EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF
without	clean		33.82	0.098	28.92	0.096	18.26	0.067	15.38	0.060
partial [18]	clean	reverberated	35.53	0.096	32.35	0.099	22.32	0.081	14.50	0.063
		noisy	29.84	0.094	18.08	0.078	22.55	0.081	11.99	0.058
full	reverberated		33.37	0.099	31.85	0.098	13.21	0.061	10.47	0.051
	noisy		18.45	0.075	14.16	0.062	13.89	0.063	11.20	0.051

Table 3: *Speaker recognition results using clean enrollment data and noisy or enhanced test data.*

Type of MCT	Training of UBM and T matrix	Training of LDA and PLDA	Clean enrollment, noisy test data				Clean enrollment, enhanced test data			
			LDA		PLDA		LDA		PLDA	
			EER	mDCF	EER	mDCF	EER	mDCF	EER	mDCF
without	clean		19.76	0.078	17.90	0.079	13.66	0.064	12.79	0.061
partial [18]	clean	reverberated	38.10	0.099	20.53	0.081	35.94	0.098	12.07	0.058
		noisy	38.72	0.100	18.47	0.075	34.87	0.097	13.74	0.067
full	reverberated		19.59	0.078	17.31	0.074	12.52	0.055	10.81	0.051
	noisy		16.85	0.069	14.57	0.066	15.00	0.064	13.66	0.064

Unlike [18], we cannot conclude that applying MCT in the enrollment stage yields the best results in all cases. This might be due to the fact that the experimental conditions in both papers are not exactly the same. In [18], the MCT dataset includes clean speech and the distortion consists of noise only (no reverberation). Furthermore, our trials involve two signals with unknown noise conditions, while [18] takes advantage of the multiple enrollment utterances provided by NIST-SRE 2012, averaging them to form the trial, thus the noisy test signal can be compared with a MCT target. However, real applications cannot frequently count on multiple signals in different noisy environments to characterize the target.

Nevertheless, MCT enrollment consistently outperforms clean enrollment when considering the best classification technique (PLDA) and the best training data (noisy training data for noisy test data and reverberated training data for enhanced test data).

4.3. Analysis of the computational cost

Previous works [18] stated that MCT applied to the UBM is time consuming. In spite of the fact that the UBM is computed a single time in an offline manner, we agree with this statement in the case when the training dataset is replicated for each condition in the test set. The training dataset then becomes huge, leading to a considerable increase of computational time for the training of the UBM and the T matrix.

In this work, we proposed to create a training dataset that spans the conditions in the test set, but without increasing the dataset size. Instead of replicating the dataset, we randomly mixed the training dataset with all the conditions, keeping the same size. Therefore the time required to train the UBM and the T matrix from multicondition data was exactly the same as the time required for clean training. The results in Tables 2 and 3 support this proposal, demonstrating that the system performance increases even when keeping the size of the training dataset and the number of UBM Gaussians unchanged.

5. Conclusions

In this paper, we proposed to apply MCT over all training stages of the i-vector PLDA framework for improving the performance of speaker recognition in noisy environments. In order to keep the same computational cost of the system as without MCT, we created the MCT dataset using several conditions randomly mixed with the training data, without increasing the dataset size. We conducted a series of experiments on reverberated and noisy speech with diverse, unknown conditions, as is the case in several real applications scenarios.

The results showed that the proposed full MCT provides significantly better EER and mDCF than clean training or partial MCT applied to the classification stage (LDA and PLDA) only. Crucially, these results were obtained using a single signal to characterize the target during enrollment. This is a more realistic setup compared to certain past evaluations, since multiple enrollment signals in different conditions are not available in most real applications.

These results, together with the fact that the computational cost does not increase with the introduction of MCT, lead to the conclusion that this is an efficient and effective way to take advantage of the MCT concept to increase the robustness of speaker recognition in noisy environments.

We also evaluated the impact of state-of-the-art multichannel speech enhancement and showed that it improves the EER and the mDCF both alone and in combination with full MCT. Even though it is frequently believed that single-channel speech enhancement methods deteriorate the speaker discriminative information in the signal, affecting the recognition accuracy, multichannel speech enhancement is worth reconsidering in that light.

6. Acknowledgements

This work has been partly realized thanks to the support of the Region Lorraine and the CPER MISN TALC project.

7. References

[1] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Support vector machine versus fast scoring in the low-dimensional total variability space for speaker verification," in *Interspeech*. Brighton, UK: ISCA, 2009, pp. 1559–1562.

[3] ——, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[4] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matějka, and N. Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *ICASSP*. Prague, Czech Republic: IEEE, 2011, pp. 4832–4835.

[5] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Wiley, 2009.

[6] L. Deng, "Front-end, back-end, and hybrid techniques for noise-robust speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*. Springer, 2011, pp. 67–99.

[7] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[8] K. S. Rao and S. Sarkar, *Robust Speaker Recognition in Noisy Environments*. Springer Science+Business Media, 2014.

[9] G. M. Davis, *Noise Reduction in Speech Applications*. New York: CRC Press LLC, 2002.

[10] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *ICASSP*. Texas, USA: IEEE, 1987, pp. 705–708.

[11] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *Spoken Language Technology Workshop*. Miami, USA: IEEE, 2012, pp. 131–136.

[12] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *ICASSP*. Vancouver, BC: IEEE, 2013, pp. 7319–7323.

[13] A. Gorin and D. Jouvet, "Structured GMM based on unsupervised clustering for recognizing adult and child speech," in *2nd International Conference on Statistical Language and Speech Processing*, Grenoble,France, 2014, pp. 108–119.

[14] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Berlin, Germany: Springer-Verlag, 2008.

[15] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Speech and Audio Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[16] D. Garcia and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*. Florence, Italy: ISCA, 2011, pp. 249–252.

[17] J. Villalba and E. Lleida, "Handling i-vectors from different recording conditions using multi-channel simplified PLDA in speaker recognition," in *ICASSP*. Vancouver, BC: IEEE, 2013, pp. 6763–6767.

[18] P. Rajan, T. Kinnunen, and V. Hautamaki, "Effect of multicondition training on i-vector PLDA configurations for speaker recognition," in *Interspeech*. Lyon, France: ISCA, 2013, pp. 3694–3697.

[19] D. Martinez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny, and E. Lleida, "Unscented transform for ivector-based noisy speaker recognition," in *ICASSP*. Florence, Italy: IEEE, 2014, pp. 4042–4046.

[20] A. El-Solh, A. A. Cuhadar, and R. A. Goubran, "Evaluation of speech enhancement techniques for speaker identification in noisy environments," in *Proc. ISMW*, 2007.

[21] NIST: National Institute of Standards and Technology. [Online]. Available: http://www.nist.gov/speech

[22] T. Rossing, Ed., *Springer handbook of acoustics*. Springer, 2007.

[23] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[24] M. McLaren and D. V. Leeuwen, "Source normalised LDA for robust speaker recognition using i-vectors from multiple speech sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 755–766, 2011.

[25] ——, "To weight or not to weight: Source normalised LDA for speaker recognition using i-vectors," in *Interspeech*. Florence, Italy: ISCA, 2011, pp. 2709–2712.

[26] ——, "Source normalised and weighted LDA for robust speaker recognition using i-vectors," in *ICASSP*. Prague, Czech Republic: IEEE, 2011, pp. 5456–5459.

[27] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *ICASSP*, Kyoto, Japan, 2012, pp. 4253–4256.

[28] M. Senoussaoui, P. Kenny, P. Dumouchel, and F. Castaldo, "Well-calibrated heavy tailed Bayesian speaker verification for microphone speech," in *ICASSP*, Prague, Czech Republic, 2011, pp. 4824–4827.

[29] K. Simonchik, T. Pekhovsky, A. Shulipa, and A. Afanasyev, "Supervized mixture of PLDA models for cross-channel speaker verification," in *Interspeech*. Portland, USA: ISCA, 2012, pp. 1684–1687.

[30] E. Vincent, J. Barker, S. Watanabe, J. L. Roux, F. Nesta, and M. Matassoni, "The second 'CHIME' speech separation and recognition challenge: Datasets, tasks and baselines," in *ICASSP*. IEEE, 2013, pp. 126–130.

[31] ——, "The second 'CHIME' speech separation and recognition challenge: An overview of challenge systems and outcomes," in *ASRU*, 2013, pp. 162–167.

[32] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118 – 1133, May 2012.

[33] Y. Salaün, E. Vincent, N. Bertin, N. Souviraà-Labastie, X. Jaureguiberry, D. Tran, and F. Bimbot, "The flexible audio source separation toolbox version 2.0," in *Show and Tell of ICASSP*. Florence, Italy: IEEE, 2014.